MASKING MICRODATA FILES

Jay J. Kim and William E. Winkler, Bureau of the Census

ABSTRACT

Government agencies collect many types of data, but due to confidentiality restrictions, use of the microdata is often limited to sworn agents working on secure computer systems at those agencies. These restrictions can severely affect public policy decisions made at one agency that has access to nonconfidential summary statistics only. This necessitates creation of microdata which not only meets the confidentiality requirements but also has sufficient utility. This paper describes a general methodology for producing public-use data files that preserves confidentiality and allows many analytical uses. The methodology masks quantitative data using an additive-noise approach and then, when necessary, employs a reidentification/swapping methodology to assure confidentiality. One of the major advantages of this masking scheme is that it also allows obtaining precise subpopulation estimates, which is not possible with other known masking schemes. In addition, if controlled distortion is applied, then a prespecified subset of subpopulation estimates from the masked file could be nearly identical to those from the unmasked file. This paper provides the theoretical underpinning of the masking methodology and the results of its actual application using examples.

KEY WORDS: Confidentiality, Noise Inoculation, Reidentification, Swapping

1. INTRODUCTION

While many types of data are collected by government agencies, use of the microdata files is often limited to sworn agents working on secure computer systems at those agencies. The confidentiality restrictions can severely affect public policy decisions made at one agency that has access to nonconfidential summary statistics but not to the microdata that are collected at two or more other agencies. The application of this paper is in producing a public-use data base that contains much data from the March Supplement to the Current Population Survey (CPS) and income data from the Internal Revenue Service (IRS) 1040 Form. The data are for use by the Department of Health and Human Services (HHS) in setting policy regarding earned income credit and other benefits. The microdata is masked in such a manner that both Bureau of the Census and IRS confidentiality restrictions are met. No masked IRS quantitative data can alone be used in reidentifications.

The main methodology is an additive-noise approach (Kim 1986) for masking multivariate normal data that preserves confidentiality and can preserve many essential characteristics of the data such as means, variances, and correlations. The CPS and IRS data of the application are known to be approximately multivariate normal. While the methodology has been extended to general data distributions (Sullivan and Fuller 1989, 1990; also Little 1993), the extension involves transforming general data to multivariate normal, masking, and then transforming the masked data back to the original scale. As we begin with multivariate normal data, we need not be concerned with the two additional transformation steps of the more general Sullivan-Fuller methods. We do note that the set of general software that we developed for arbitrary multivariate normal data could be extended to the general data by inclusion of software performing the two Sullivan-Fuller transforms.

The secondary methodology of this paper is a sophisticated reidentification/swapping technology that is based on existing record linkage concepts (Winkler 1994, 1995). The matching software uses the masked CPS and IRS quantitative data along with other variables such as age, race, sex, and State to produce reidentifications with the original merged file of unmasked CPS and IRS data. Since we know true matching status, we can minimize the number of pairs of records in which quantitative data is swapped. While swapping can help preserve confidentiality, it can reduce the analytic usefulness of the file (Little 1993). By minimizing swapping and preserving means and covariances on specified subdomains, we assure the analytical usefulness of the final file as we show later.

The outline of this paper is as follows. In the second section of this paper we describe the data files and the methodologies for additive-noise masking, reidentification/swapping, and controlled-distortion. The third section provides results. In the fourth section, we describe how the methods of this paper can be used to verify the analytic validity of public-use files that are produced, discuss some of the limitations of the masking methodology, and provide an overview of the general software we developed. The final section consists of summary and conclusions.

2. **DATA AND METHODS**

This section describes the data, the masking methodology, the reidentification/swapping methodology, and the controlled-distortion methodology.

2.1. Data to be Masked

The original unmasked file of 59,315 records is obtained by matching IRS income data to a file of the 1991 March CPS data. The fields from the matched file originating in the IRS file are as follows:

I) Total income;
ii) Adjusted gross income;
iii) Wage and salary income;
iv) Taxable interest income;
v) Dividend income;
vi) Rental income;
vii) Nontaxable interest income;
viii) Social security income;

ix) Return type;
x) Number of child exemptions;
xi) Number of total exemptions;
xii) Aged exemption flag;
xiii) Schedule D flag;
xiv) Schedule E flag;
xv) Schedule C flag; and
xvi) Schedule F flag.

The file also has a match code and a variety of identifiers and data from the public-use CPS file. Because CPS quantitative data are already masked, we do not need to mask them. We do need to assure that the IRS quantitative data is sufficiently well masked so that it cannot easily be used in reidentifications, either by itself or when used with identifiers such as age, race, and sex that are not masked in the CPS file. Because the CPS file consists of a 1/1600 sample of the population, it is easier to minimize the chance of reidentification. We primarily need be concerned with higher income individuals or those with distinct characteristics that might be easily identified even when sampling rates are low.

2.2. Masking Methodology

Masking is via an additive noise approach (Kim 1986, see also Sullivan and Fuller 1989, Sullivan and Fuller 1990, and Little 1993). Adding random noise with the same correlation structure as the original unmasked data is currently the only method (Little 1993) that preserves correlations. Appendix A.3 allows us to determine means and covariances on arbitrary subdomains. Theoretical details are in Appendixes A.1, A.2, A.3, and A.4. Masking is done according to the following steps:

i) Calculate the variance/covariance for income types iii) through viii) in subsection 2.1. This results in a 6×6 variance/covariance matrix.

ii) Take c×100 percent of the above variance/covariance and generate random numbers using subroutine *RNMVN* in International Mathematical and Statistical Library (IMSL). Note that *RNMVN* requires the users to provide the variance/covariance which the generated random numbers should have.

This process produces 59,315×6 matrix of random numbers. The expected value of the generated random numbers for each of the 6 arrays is 0.

iii) Add the random numbers generated in ii) to the income fields in section 2.1. Note that both the raw income data in section 2.1 [income types iii) through viii)] and the noise in step ii) of this section are of matrix 59,315×6. Thus the addition is elementwise over the matrices.

iv) Sum up incomes for each individual for income types iii) through viii) in section 2.1 and calculate the difference between the sum and the total income, and the difference between the sum and the adjusted gross income.

v) Sum up noise inoculated incomes of types iii) through viii) for each individual. Add to the sum of the perturbed incomes the difference between the sum of raw incomes and the total income calculated in step iv) above.

This gives the masked total income. Masked adjusted gross income is produced similarly.

Six income variables are masked directly and the remaining two are masked in a manner that preserves sums. If top-coding is required for the incomes at, say, 200,000 (or -200,000), it can be done after the above five steps. In some situations, data providers censor outliers prior to masking because outliers (even when masked) are particularly easy to reidentify. In our approach, we specifically assume that data are not censored because censoring reduces the analytic validity of the masked file. A masked file is *analytically valid* if, for a (set of) analysis(es), it will give approximately the same numbers and yield the same conclusions as the unmasked (original true) file. The subdomain adjustment formulas (Appendix A.3) assure that subdomain analyses with the masked data are analytically valid because means and covariances are preserved. When we refer to accuracy as being good, we mean that estimates in the masked or masked/swapped data are quite consistent with estimates in the unmasked data. It is straightforward to make modifications to deal with censored data.

As the users might want to tabulate the counts of individuals depending on the recipiency status of various IRS income and the noise inoculation completely changed the zeros and non-zeros both alike, we add flags indicating whether each amount of unmasked income was zero or not. This allows them to analyze the data for recipient group and nonrecipient group, separately.

Even after masking, it may be possible to reidentify a certain proportion of records in the masked file with the original, corresponding records in the unmasked file. While the 1/1600 sample assures that most mid-to-low income individuals cannot be reidentified in the entire population using information from the public-use file, some individuals with high incomes or unusual combinations of age, sex, race and income characteristics might be reidentified. Specifically, if we can reidentify a mid-income record across masked and unmasked sample files and there are 2000 individuals in the population with essentially the same characteristics as those that were used in the reidentification, then there is only a 1 in 2000 chance of a reidentification. In other words, it is not possible to reidentify such a mid-income individual in the entire population via information in the public-use file. However, it may still be possible to reidentify individuals with high incomes or with unusual characteristics. To minimize the chance of reidentification, we need to employ additional procedures in a manner that does not eliminate the analytical usefulness of the public-use file. Such minimization may be possible because we are the data providers and have knowledge of the exact truth of reidentifications between unmasked and masked sample files.

2.3. Reidentification/Swapping Methodology

To determine how much reidentification is possible, we proceed in two stages. First, we match the merged raw data file against the masked file using record linkage software (Winkler 1994). Based on the reidentification rate, we next swap quantitative data according to a proportion that minimizes the chance of reidentification.

During the first stage, we use blocking variables such as age, race, sex, and State code and matching variables such as the IRS income and CPS quantitative variables. Blocking is a record linkage term that means that we only consider pairs that agree exactly on the blocking variables. The quantitative matching variables need not agree exactly. String comparators and other advanced metrics are used in computing distances between records in a manner that is compatible with the main decision rule. The matching decision rule is based on an information-theoretic extension of the likelihood ratio test (Fellegi and Sunter 1969) that assigns scores to each pair based on a function of their associated likelihood ratios. Likely reidentifications, called matches, are given higher scores, and other pairs, called nonmatches, are given lower scores. To best separate the pairs into matches and nonmatches, we use a version of the EM algorithm for latent classes (Winkler 1994) that determines the best set of matching parameters under certain model assumptions which are not seriously violated in this particular situation. To force 1-1 matching efficiently, we apply an assignment algorithm due to (Winkler 1994). When a few matching pairs in a block can be reasonably identified, many other pairs can be easily identified via the assignment algorithm. The assignment algorithm has the effect of drastically improving matching efficacy, particularly in reidentification experiments of the type given in this paper.

During the second stage, we first collapse cells (age × race × sex) to assure that there are sufficient candidates for swapping. The collapsing strategy is similar to those used in sampling and nonresponse imputation. Within collapsed cells we randomly swap quantitative data according to a proportion that we specify. Since we know true matching status, we can minimize the swapping proportion because we know exactly which pairs are reidentifications. We note that the specific set of reidentifications varies with each different seed value used at the masking stage. Swapping preserves means and correlations in the subdomains on which it was done and in unions of those subdomains. On arbitrary subdomains, however, collapsing and the amount of swapping can adversely affect the analytic validity of the files. If swapping is done such that each record that is swapped is only swapped with another record in that subdomain, then we say that we have *controlled* that subdomain. Means and correlations among swapped variables within controlled subdomains are necessarily the same. We cannot hope for confidentiality while providing analytic validity in arbitrary subdomains above a certain size. If we were to provide such analytic validity in subdomains above a certain size, then we would necessarily be able to reidentify every record in the file.

We say that a specified record has *disclosure risk* of x percent if the estimated probability of the match being correct is x percent. Belin and Rubin (1995) have given a method of estimating the probability of a match being correct that requires a training set and does not work with the data of this paper. An alternative method of Winkler (1994), which requires an ad hoc intervention and no training set, is used to estimate disclosure risk.

2.4. Controlled Distortion

In this section, we introduce a third procedure, called *controlled distortion,* provide a justification for using it, and relate it to the noise addition and swapping procedures of the two previous sections. Addition of noise has the advantages that we know the distribution of the noise that is added to each record and that we can deduce the nonmasked means and variances in arbitrary subdomains via a procedure in Appendix A.3. The main disadvantage of noise addition is that individual records with quantitative data that is significantly different from other records are *easily identifiable (ei)*. An *ei*-record is one whose masked data can still be used to match it against the correctly corresponding unmasked data record. The first way of dealing with *ei*-records is data swapping. Within a subdomain defined by records agreeing on characteristics such as age range, sex, and race, we can swap all (or an arbitrary subset) of an *ei*-record's quantitative data with the corresponding quantitative data from another record in the subdomain. The swapping can be against a random record or against the second best match. The best match is the *ei*-record. Swapping has the advantage that it is straightforward in concept. If only a small proportion of records is swapped, then means and correlations may not be seriously distorted.

The disadvantage of swapping is that means and correlations are only preserved for the subdomains in which swapping is done. For arbitrary subdomains, means and correlations for masked data may exhibit large deviations from the means and correlations for unmasked data. We can partially address the large-deviation problem as follows. During the first swapping pass, identify the *ei*-records whose second best matches are not close and do not swap them. Enlarge the subdomains to assure that each remaining, unswapped *ei*-record can be matched against a record whose quantitative data is much closer. Perform swapping in the larger subdomains. The advantage of the two-pass procedure is that it will (nearly) preserve means across arbitrary subdomains. The deviations of correlations, however, may not be as well preserved.

Controlled distortion is a procedure on a subdomain A where we change the values in one record arbitrarily and also perform a series of complementary changes so that means and covariances are preserved in the subdomain. In Appendix A.5, we show that valid controlled distortion procedures exist provided that the subdomain contains at least L^2 records where L is the number of variables for which we preserve means and covariances. The advantage of controlled distortion is that *ei*-records can be distorted in an arbitrary manner specified by the data provider and can assure confidentiality. Controlled distortion has the same disadvantage as swapping because means and correlations cannot be preserved across arbitrary subdomains.

3. **RESULTS**

In this section we begin with results for the files in which masking and no swapping have taken place. This allows us to show how the additive-noise approach yields files having means and covariances nearly identical to the original, unmasked file on many subdomains. We then present results for the files in which both masking and swapping have been performed. We conclude with results on disclosure risk in the files.

3.1. Utility of the Full Sample Data

Since the model building requires mean and variance/covariance or correlation of the variables involved, statistics were calculated for six variables in the raw and masked data. The means of the raw and masked data are almost identical (Table 1).

Table 1. Means of Raw and Masked Data

Type	Raw	Masked
Wage	23,799	23,784
Tax Int	1,825	1,823
Div	587	587
Rent	1,190	1,187
Ntax Int	342	342
Soc Sec	947	948

Table 2. Correlation for Raw and Masked Data

	Raw	Masked
Wage vs Dividend	.18	.18
Wage vs Tax Int	.12	.12
Dividend vs SS	.12	.12
Tax Int vs Rent	.08	.08
Dividend vs Rent	.04	.04
Ntax Int vs SS	.04	.04

Table 2 shows that all correlations are the same to two decimal places. As indicated earlier, total and adjusted gross income were masked indirectly by summing up masked components of the total and adjusted gross income except the difference between the sum of the unmasked data and total or adjusted gross income.

The means of the total and adjusted gross income from the masked data are virtually identical as those from the unmasked data. They differ by less than 0.0007. This can be expected since the noise was added to all components has zero expected value and the sample size is quite large. Similarly, the variance of the total and adjusted gross income from the masked data are virtually identical to those from the unmasked data.

3.2. Subdomain Estimation - before Swapping or When Swapping Was Controlled for Subdomain

In this subsection we examine subdomain estimation which is of special interest to data users. Appropriate subdomain estimation formulas for the masked data are given in Appendix A.3. Subdomain means are not affected by the masking. Only a minor adjustment is needed to the variance/covariance according to the formula shown in the appendix because the amount of noise added is low (in terms of the variance/covariance). The adjustment also has almost no effect on the correlation.

To determine how well the subdomain adjustment formulas work, we compute estimates for those persons whose "return type" is 4, (unmarried head of household return). Generally, the estimates of means from the masked data are excellent. For five items, they are virtually identical with those from the unmasked data. However, the estimate of mean nontaxable interest (61) from the masked data is more than 10 percent off from the mean (70) of the unmasked data. Tables 3 shows correlations between the income variables for the unmasked and masked data, respectively.

Table 3. Correlation for Raw and Masked Data for Return Type = 4

	Raw	Masked
Wage vs Dividend	.027	.029
Wage vs Tax Int	.108	.105
Dividend vs SS	.155	.154
Tax Int vs Rent	.172	.171
Dividend vs Rent	.040	.039
Ntax Int vs SS	.056	.052

Estimated correlations of the masked data on this subdomain are generally good, agreeing with the unmasked data to two decimal places. While we do not show it here, the same statistics were estimated from the masked data for other subdomains: Return type=1 (single return) and Schedule C=1 (Schedule C was filed in the tax return). Similar close agreements were found.

Thus far we have observed the behavior of subdomain estimates when the subdomain is formed by a variable which is not masked. What happens when the subgroup is formed by a masked variable itself? By adding noise, in effect we expand the range of values the variable can take. If we use the same cutoff to form a subgroup for both the unmasked and masked data, there is no guarantee that the same elements will be in the same group in both data sets. To check on the performance of statistics when the subdomain is formed based on the masked variable, wage and salary, shortened to wage, is chosen to be used as a classification variable. The subdomain consisted of records having wage less than 15,000. The subdomain in the unmasked file had 28,268 records and the comparable subdomain in the masked file had 28 more records. Means were virtually identical. Correlations were virtually identical, differing only in the third decimal place.

3.3. Subdomain Estimation - When Swapping Was Not Controlled on the Subdomain

Our swapping procedure involved swapping only the eight IRS income fields and three CPS income fields such as wage (it will be called CPS Wage), adjusted gross income (it will be called CPS Agi) and aggregated sum of rent (net rent), dividend and interest (it will be called CPS Prop). This swapping procedure will not generally preserve means and covariances on arbitrary subdomains such as the subdomain determined by those records corresponding to a person having filled out a Schedule C return. Table 4 compares means for two swapping rates, 5% and 20% with the raw and unmasked data.

Table 4. Means before And after Swapping for Schedule C Users, n = 7,819

	Raw	Masked	5% Swap	20% Swap
Wage	24,715	24,677	25,338	26,891
Rent	2,820	2,822	2,779	2,746
Tax Int	2,178	2,174	2,171	2,145
Dividend	783	779	773	755
Ntax Int	393	391	366	346
Soc Sec	790	790	803	822

The table shows that I) subdomains estimates using raw and masked data agree closely; ii) subdomain estimates on the 5-percent swapping file still agree closely with raw and masked data estimates; and iii) 20-percent swapping differ by a greater amount from the raw and masked data estimates than 5-percent estimates.

The next table shows some selected correlations.

Table 5. Correlations before and after Swapping for Schedule C Users, n = 7,819

			Swap Rate	
Fields	Raw	Masked	5%	20%
Wage, Dividend	.6361	.6352	.6143	.6217
Wage, Tax Int	.1903	.1900	.2425	.2413
Dividend, SS	.1535	.1547	.1528	.1346
Tax Int, Rent	.1984	.1978	.1967	.2167
Dividend, Rent	.1291	.1285	.1265	.1304
Ntax Int, SS	.1057	.1062	.1181	.0957

Swapping has some impact on the correlations but still yields correlations that are good. Accuracy is better with 5-percent swapping than with 20-percent swapping.

3.4. Reidentifcation and Confidentiality

We investigated the masked file and the masked/swapped file. The risk of disclosure for the masked file is somewhat high. As much as 0.8% of the records have a probability of disclosure above 20%; the remaining 99% have a disclosure risk of less than 0.02%. The disclosure risk for all records in the masked/swapped file is below 0.1%.

4. **DISCUSSION**

The discussion covers how representative the masking procedures are, their ability to produce analytically valid files, and some of their limitations. The section also provides an overview of our general computer software for masking arbitrary multivariate normal files.

4.1. Representativeness of Results

The masking/swapping procedures were repeated with two additional seed numbers for the random noise-generation routine. The correspondences of means and correlations between unmasked and masked/swapped files were consistent with those given in this paper. We note the actual set of reidentification/swaps varies with the seed numbers because reidentifications depend on how close individual masked data records are to corresponding unmasked data records. The closeness is dependent on the random noise which varies with the seeds.

4.2. Analytic Validity of Public-Use Files

Swapping can distort the correlations, particularly on subdomains. We suggest releasing two copies (one for each seed used in the random number generator) of the masked/swapped files. If users cannot reproduce a statistical analysis using data from one copy that was done on the other copy, then they can be assured that the public-use file will not support the attempted analysis. In that case, there are two recourses. The first is for the data providers to supply two more copies of the public-use file that have been masked and swapped in a manner that supports the originally attempted analysis. If that is not possible, then the only second recourse is to have the statistical analysis performed on the original, unmasked data.

4.3. Limitations

When a masked/swapped continuous variable is used for categorization, the number of observations in categories may not be close to those from the unmasked data. This is because the categorization implicitly corresponds to subdomains in which swapping may not be controlled. The summary statistics for categories between unmasked and masked/swapped data can be consistent if the sizes of the categories are large. If the subdomain of interest is of small size, then we should be careful about using statistics for the subdomain.

4.4. Software

The current version of the computer software can be used for masking and swapping general multivariate normal files. The first program (in SAS) produces an output file consisting of the variance/covariance matrix for the raw data. The second program (in FORTRAN) calls the IMSL routine *RNMVN* to produce random multivariate noise with the same variance/covariance as the raw data. The third combines raw data and noise to produce the masked file. The fourth program (in C) does swapping. All software is portable provided the IMSL routine *RNMVN* is available. If *RMNVN* is not available, then similar types of multivariate normal noise can be generated using SAS or various public-domain random number generation packages.

5. **SUMMARY AND CONCLUSIONS**

We demonstrated a methodology for producing a confidential, analytically valid, public-use file that contains eight income fields from the 1990 IRS Tax Return file and the remaining data from the 1991 CPS public-use file. The file was produced in two stages. The first stage consisted of adding random noise with the same correlation structure as the original, unmasked data. The second stage involved reidentifying and swapping records via a record linkage approach.

Whereas the masked file is analytically valid with means and correlations (even in many subdomains) that are very close (3 decimal places) to means and variances in unmasked files, the risk of disclosure for the masked file is somewhat high. As much as 0.8% of the records have a probability of disclosure above 20% because they have unusual combinations of characteristics that make it relatively straightforward to distinguish from other recrods. The remaining 99% have a disclosure risk of less than 0.02%. The reidentification/swapping procedure reduced the disclosure risk in the masked/swapped file to below 0.1% while preserving means and covariances in a specified set of subdomains. For the entire domain, means and correlations from the masked/swapped file were typically within 3 decimal places from the corresponding means and correlations in the unmasked file. Deviations in many subdomains were higher; sometimes deviating in the second decimal place.

REFERENCES

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association*, **90**, 694-707.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 303-308.

Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods,* 456-461.

Little, R. J. A., (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9**, 407-426.

Sullivan, G., and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 802-807.

Sullivan, G., and Fuller, W. A. (1990), "Construction of Masking Error for Categorical Variables," American Statistical Association, *Proceedings of the Section on Survey Research Methods,* 435-439.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage, American Statistical Association, *Proceedings of the Section on Survey Research Methods,* 467-472.

Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox (ed.) *Business Survey Methods,* New York: J. Wiley, 355-384.

APPENDIX A.1. NOISE INOCULATION

Let x_{ij} be the unmasked jth income value of the ith person, I=1,,,59,315 and j=3,,,,8. Also let e_{ij} be the noise added to x_{ij} and $y_{ij} = x_{ij} + e_{ij}$, I=1,,,59,315 and j=3,,,,8.

Let X be the matrix having x_{ij} as elements, I=1,2,,,,59,315 and j=3,4,,,8. Similarly, E = {e_{ij}} and Y = {y_{ij}}. Let Var(X) = E. Then we are using Exp(E) = $\underline{0}$ and Var(E) = cE, where $\underline{0}$ is a 59,315x6 matrix having all 0 elements. Thus E(Y) = E(X) and Var(Y) = (1+c)E. The variance of unmasked variables can be recovered by {1/(1+c)}Var(Y).

Let $x_{i.} = \overset{8}{\underset{j=3}{E}} x_{ij}$, I=1,2,,,59,315 and $diff_{ij} = x_{ij} - x_{i.}$, I=1,2,,,59,315 and j=1,2. The masked total and adjusted income can be expressed as follows.

$y_{i1} = \overset{8}{\underset{j=3}{E}} y_{ij}$ + $diff_{i1}$, I=1,2,,,,,59,315

and

$y_{i2} = \overset{8}{\underset{j=3}{E}} y_{ij}$ + $diff_{i2}$, I=1,2,,,,,59,315.

APPENDIX A.2. DERIVATION OF VARIANCE/COVARIANCE

Since unmasked income and noise are independent, Var(y) = (1+c)F^2 for each component income, where F^2 is the variance of x. Total income can be reexpressed for deriving variance and covariance.

$y_{i1} = x_{i1} + \overset{8}{\underset{j=3}{E}} e_{ij}$.

Thus Var(y_1) = Var(x_1) + Var($\overset{8}{\underset{j=3}{E}} e_j$), where y_1 and x_1 are masked and unmasked total income (first IRS income variable on the file), and e_j is the noise added to the jth income disregarding subscript for person number, which can be reexpressed as follows:

$$Var(y_1) = Var(x_1) + \sum_{j=3}^{8} cVar(x_j).$$

Covariance between total income and each component of the total income can be expressed as follows:

$$Cov(y_1,y_j) = Cov(x_1+ \overset{8}{\underset{j=3}{E}} e_j, x_j+e_j) = Cov(x_1, x_j) + Var(e_j) + \sum_{i \neq j} Cov(e_i,e_j)$$

The variance of adjusted gross income and covariance between the adjusted gross income and each of the components of the adjusted gross income can be derived similarly. However, the covariance between total income and adjusted gross income is a little different from what we have seen.

$\text{Cov}(y_1,y_2) = \text{Cov}(x_1+ \overset{8}{\underset{j=3}{\text{E}}} e_j, + x_2+ \overset{8}{\underset{j=3}{\text{E}}} e_j) = \text{Cov}(x_1, x_2) + \text{Var}(\overset{8}{\underset{j=3}{\text{E}}} e_j) =$ $\text{Cov}(x_1,x_2) + \overset{8}{\underset{j=3}{\text{E}}} Var(e_j)$.

The covariance between a masked variable and an unmasked variable is the same as that between two unmasked variables, i.e.,
$\text{Cov}(y_i,x_j) = \text{Cov}(x_i,x_j)$.

APPENDIX A.3. SUBDOMAIN ESTIMATION

Let s stands for a subdomain, i.e., x_j^s and y_j^s are the unmasked (x) and masked variable (y) defined for subdomain s for variable j and $\text{F}_j^{2^s}$ is the variance of x_j^s.

Since the noise is generated for the full data set (rather than for each subdomain), the relationship between x_j^s and y_j^s are as follows:

$$y_j^s = x_j^s + e_j.$$

Since $\text{E}(e_j) = 0$, $\text{E}(y_j^s) = \text{E}(x_j^s)$.
Also since $\text{Var}(e_j) = c\text{F}_j^2$ and $\text{Var}(y_j^s) = \text{Var}(x_j^s) + c\text{Var}(x_j)$,

$\text{Var}(x_j^s) = \text{F}_j^{2^s} = \text{Var}(y_j^s) - c\text{F}_j^2$.

Since $\text{Var}(y_j) = (1+c)\,\text{F}_j^2$ and $\text{F}_j^2 = \text{Var}(y_j)/(1+c)$, to recover the variance of the unmasked variable from the masked data we can use

$$\text{F}_j^{2^s} = \text{Var}(y_j^s) - \frac{c}{1+c}\text{Var}(y). \qquad (1)$$

Note the above formula for variance of a unmasked variable for a subdomain is a linear combination of variance of the masked variable for the subdomain and a fraction of variance of the masked variable for the full data set.

The covariance between two masked variables for a subdomain, y_j^s and y_k^s, can be derived similarly. Note that

$$\text{Cov}(y_j^s,y_k^s) = \text{E}[(x_j^s+e_j),(x_k^s+e_k) - \text{E}(x_j^s+e_j)\text{E}(x_k^s+e_k). \qquad (2)$$

Since we generate the random noise such that the noise independent of the unmasked variable and $\text{Cov}(e_j,e_k)=c\text{Cov}(x_j,x_k)$, equation (2) can be reduced to $\text{Cov}(x_j^s,x_k^s) + c\text{Cov}(x_j,x_k)$. Thus,

$$\text{Cov}(x_j^s,x_k^s) = \text{Cov}(y_j^s,y_k^s) - c\text{Cov}(x_j,x_k).$$

But as before, $\text{Cov}(x_j,x_k) = \text{Cov}(y_j,y_k)/(1+c)$. Thus to recover the covariance between the unmasked variables from the masked data, we can use

$$\mathrm{Cov}(x_j^s, x_k^s) = \mathrm{Cov}(y_j^s, y_k^s) - \frac{c}{1+c}\mathrm{Cov}(y_j, y_k). \qquad (3)$$

Note the semblance of equation (3) with equation (1).

If one variable is masked but the other is not, then the covariance between the variables is the same as the covariance between two unmasked variables, i.e.,

$$\mathrm{Cov}(x_j^s, y_k^s) = \mathrm{Cov}(x_j^s, x_k^s).$$

This is because the noise is generated independent of the unmasked variable.

APPENDIX A.4. PARAMETER ESTIMATION FROM THE MASKED DATA

As mentioned before E(Y) = E(X) and Var(Y) = (1+c)E. Thus variance of unmasked variables can be recovered by dividing Var(Y) by (1+c). The same holds true for the sample estimates, i.e., the estimated variance of unmasked variables can be recovered by dividing by (1+c) the estimated variance of masked variables. However, correlation is not affected by the added noise. Note in masking both variance and covariance are inflated by (1+c).

As c is increased, deviation of estimates of the masked data could increase from those of the unmasked, which is affected by amplified deviation.

There is no random number generator which can generate random numbers that produce the specified mean and variance because of the random variability of noise. Thus if we calculate mean and variance after adding noise, it is mostly likely that sample mean $\overline{y}$ will be different from $\overline{x}$, and the variance of y will be different from (1+c) times the variance of x.

APPENDIX A.5. CONTROLLED DISTORTION

This appendix provides the proof that controlled distortion is valid (i.e., can be defined so that it preserves means and correlations). We assume that the subdomain A contains L^2 records. We let (x_{i1}, x_{i2} ..., x_{iL}), I = 1, ..., L, and (y_{i1}, y_{i2}, ..., y_{iL}), I = 1, ..., L, represent original data records and controlled-distorted data records, respectively.

Theorem A.5. Let A be an arbitrary subdomain containing L^2 records where L is the number of fields. Then a valid controlled distortion procedure can be defined.

Proof. The proof is via an inductive procedure that provides the algorithm needed for the computer software. We first observe that it is sufficient to find y's such that

$$\textstyle\sum_{I=1}^{1M} x_{ij} = \sum_{I=1}^{1M} y_{ij} \text{, for } j = 1, ..., L, \qquad (1)$$

$$\textstyle\sum_{I=1}^{1M} x_{ik}\, x_{jk} = \sum_{I=1}^{1M} y_{ik}\, y_{jk}\text{, for } I \neq j = 1, ..., L, \qquad (2)$$

where we define $M = L^2$. The proof proceeds in steps. At each step, we successively consider pairs of variables. On the first step, we consider only the first two fields and the first two records. For j > 2, we take

$y_{ij} = x_{ij}$ for I = 1 and 2

and for I > 2, we take

$y_{ij} = x_{ij}$ for $j > 2$.

Thus, the components of equations (1) and (2) associated with the first two fields and the first two records reduce to

$$x_{11} + x_{21} = y_{11} + y_{21}, \tag{3}$$

$$x_{12} + x_{22} = y_{12} + y_{22}, \text{ and} \tag{4}$$

$$x_{11}\, x_{12} + x_{21}\, x_{22} = y_{11}\, y_{12} + y_{21}\, y_{22}. \tag{5}$$

Equations (3) and (4) are the means and equation (5) is the covariance. At the first step, we have no auxiliary restraints and we can distort x_{11} arbitrarily to y_{11} where we assume that $x_{11} > y_{11}$. Then, by equation (3), $x_{21} < y_{21}$. Let x_{12} and x_{22} be fixed. The equations (4) and (5), are two equations in two unknowns y_{12} and y_{22} which we can solve. For instance,

$$y_{i2} = (1/(y_{21} - 1))\,(x_{11}\, x_{12} + x_{21}\, x_{22} + y_{i1}\, x_{12} + y_{i1}\, x_{22}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \tag{6}$$

We need the additional minor restriction that $y_{21} \neq 1$. Observe that the means of all variables and the covariances between the first and second variables are preserved by the above procedure, that we have used the first two records, and that the y terms associated with the first two fields in all records beyond the first two records agree with the original x terms. We next wish to find complementary adjustments to the third field such that all means and the covariances between the first and second fields and the first and third fields are preserved.

We chose the next three records and consider the following equations:

$$y_{33} + y_{43} + y_{53} = c_1, \tag{7}$$

$$y_{33}\, y_{31} + y_{43}\, y_{41} + y_{53}\, y_{51} = c_2, \tag{8}$$

$$y_{33}\, y_{32} + y_{43}\, y_{42} + y_{53}\, y_{52} = c_3, \tag{9}$$

where $c_1 = x_{33} + x_{43} + x_{53}$, $c_2 = x_{33}\, x_{31} + x_{43}\, x_{41} + x_{53}\, x_{51} + d_2$, $c_3 = x_{33}\, x_{32} + x_{43}\, x_{42} + x_{53}\, x_{52} + d_3$, $d_2 = x_{11}\, x_{13} + x_{21}\, x_{23} - y_{11}\, y_{13} + y_{21}\, y_{23}$, and $d_3 = x_{12}\, x_{13} + x_{22}\, x_{23} - y_{12}\, y_{13} + y_{22}\, y_{23}$. We fix all y terms associated with fields 1 and 2 and with the first two records. Then equations (7), (8), and (9) are three equations in three unknowns which is uniquely solvable if the array [1 1 1, y_{31} y_{41} y_{51}, y_{32} y_{42} y_{52}] is nonsingular. We note that the values in the array and the terms c_1, c_2, and c_3 are constant. If the array is not nonsingular or we wish additional flexibility in solving for the y terms associated with the third field for the newly added records, we can add more records. This increases the number of unknowns while the number of restraints stays constant. The terms d_2 and d_3 are adjustments for the effects of the new y terms associated with the first two records for the first two fields at the previous step of the induction. At the end of the second step, we have used n_2 records, means of all fields and the covariances among the first three fields are preserved, and the y terms corresponding to records beyond record n_2 agree with the x terms for the first three fields.

At step M-1, we have used $n_{M-1} =_{def} S$ fields, all means and the covariances among the first M-1 fields are preserved. We take n_{M-1} additional records and consider the equations

$$y_{S,M} + y_{S+1,M} + ... + y_{S+M-1,M} = c_{M,1}, \tag{10}$$

$$y_{S,M}\, y_{S,1} + y_{S+1,M}\, y_{S+1,1} + ... + y_{S+M-1,M}\, y_{S+M-1,M} = c_{M,2}, \tag{10+1}$$

$$y_{S,M}\, y_{S,M-1} + y_{S+1,M}\, y_{S+1,M-1} + ... + y_{S+M-1,M}\, y_{S+M-1,M} = c_{M,M-1}, \tag{10+M-1}$$

The constants $c_{M,1}$, $c_{M,2}$, ..., and $c_{M-1,M}$ contain adjustments for the first M-2 steps that assure that the covariances between field M and the first M-1 fields are preserved. The coefficients in equations (10), (10+1), ..., and (10+M-1) are fixed because they are based on the values determined during previous steps in the induction. Also, we can take more than the minimum M records to assure that a solution to equations (10), (10+1), ..., and (10+M-1) can be obtained or that we can choose among possible solutions. Because the equations (10), (10+1), ..., and (10+M-1) can be solved, the induction is complete. ■